

Whither Come the Words?

Dr. Elizabeth D. Liddy

**Center for Natural Language Processing
School of Information Studies
Syracuse University**

A Continuum from Human to Statistical Indexing

- **Manual**
 - Controlled vocabularies
- **Mixed Initiative**
 - Machine-aided / Human-assisted
 - Machine Learning
- **Automatic**
 - Statistical indexing
 - Natural Language Processing indexing

Basic Premise

- The quality of the representation of documents determines:
 - the ‘richness’ of the indexing
 - the ‘quality’ of access to relevant information
 - the ‘value-add’ analytics the system can accomplish for users

Central Problem of IR

How to represent documents for retrieval (Blair, 1990)

- key issue in controlled vocabulary representation & searching
- still true with full-text indexing and free-text querying systems
- because documents & queries are expressed in language
 - language is complex and ambiguous
 - methods for solving the language issue are difficult
 - some IR systems don't even attempt to deal
 - major challenge of high quality information access

1. Identify indexable / queryable elements:

What is a term?

- Alpha-numeric characters between blank spaces or punctuation?
 - What about non-compositional phrases?
 - Multi-word proper names?
 - What about inter-word symbols such as hyphens or apostrophes?
 - “small business men” vs. “small-business men”

2. Represent the *concept* behind the *term*

- **Ability to take ‘terms’, and:**
 - Standardize
 - Expand to alternative ‘terms’
 - Disambiguate
- **So that the concept behind the ‘term’ is represented in both documents & queries**

Term Expansion:

Goal - add all variant terms which refer to the same concept:

- either synonymous expressions or associated terms
- use either thesaurus, semantic network, or statistically determined co-occurring terms/phrases
- inspired by success of humanly-consulted IR thesauri used in earliest systems
- relieves the user from needing to generate all conceptual variants

Term expansion:

– **Multiple approaches:**

- Knowledge-based
- Linguistic
- Statistical

Knowledge-based Thesauri

- **I. R. - style**
 - intended for human indexers and searchers
 - manually constructed for a specific domain
- **Contain synonymous, more general, and more specific terms**
 - *Use For*
 - *Broader*
 - *Narrower*
 - *Related*
- **Current question is how to utilize them appropriately in Web-based systems**

Knowledge-based Thesauri

DATABASE MANAGEMENT SYSTEMS

UF databases

NT relational databases

BT file organization

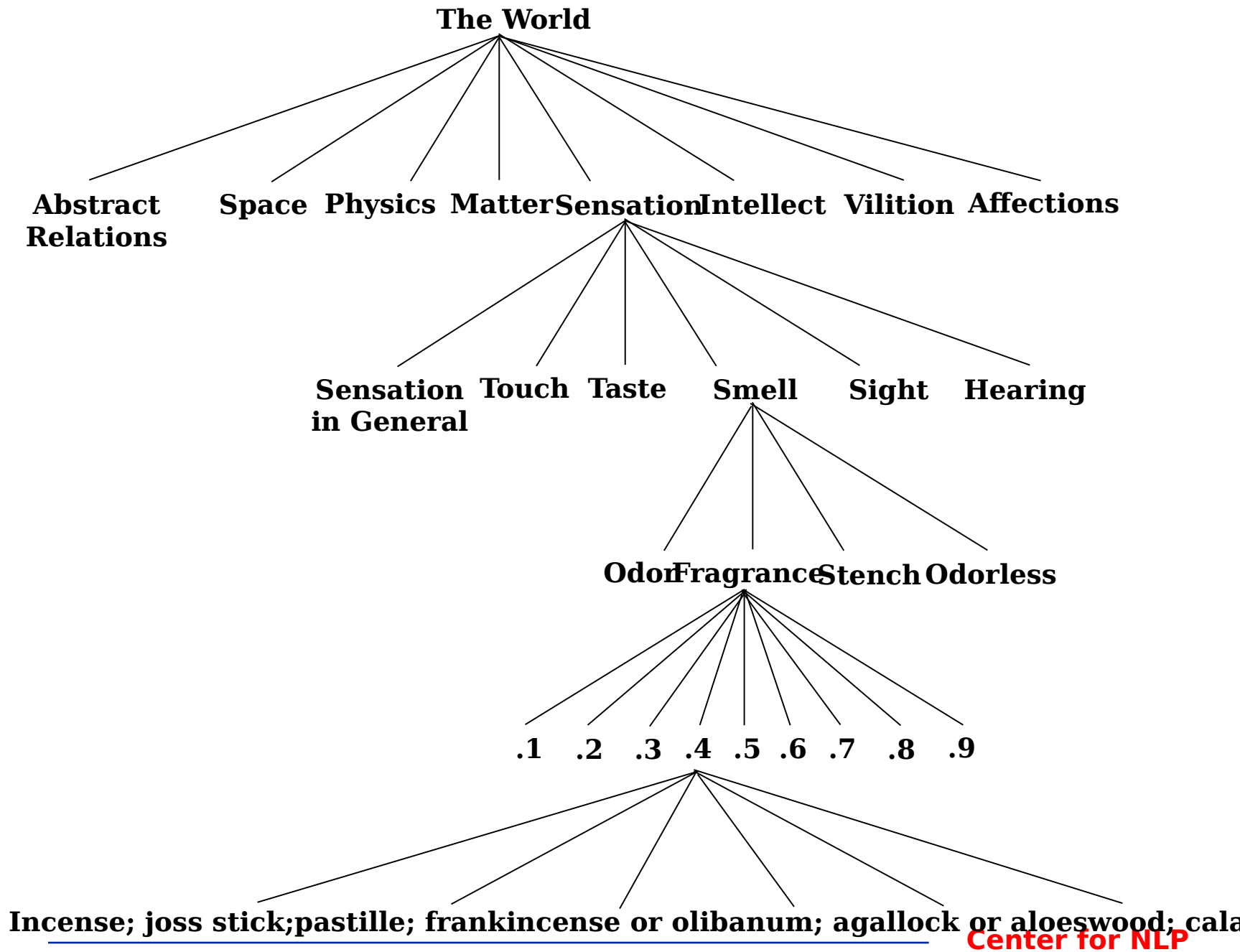
management information systems

RT database theory

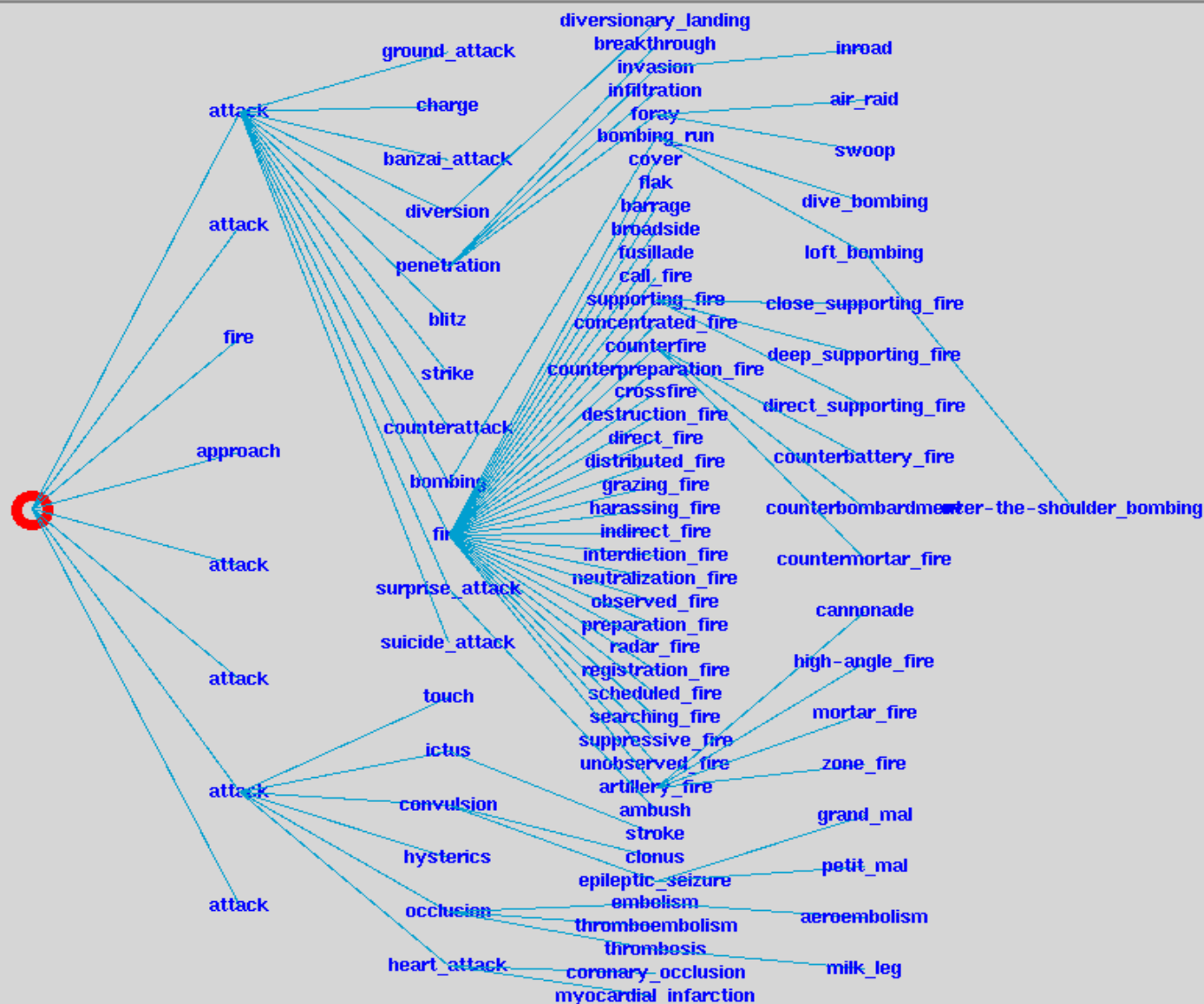
decision support systems

Linguistic Thesauri

- **General purpose style**
 - e. g. Roget's, Word Net
 - contain explicit concept hierarchies of up to 8 increasingly specified levels
- **Based on assumption that the words in a semi-colon group (RIT) or a synset (WordNet) are synonymous or near-synonymous**
 - issue / difficulty is selecting correct sense for terms



Target Word: Relation: Part of Speech: Class:



Linguistic Thesaurus Use in I R

- **Can be used on either / both documents or queries**
 - more commonly done on queries
- **Terms are expanded by adding one or all of:**
 - synonyms
 - hyponyms
 - hypernyms
- **Issues caused by:**
 - idiomatic, specialized terms
 - non-compositional phrases not in thesaurus

Process used by Voorhees '93 Research

- **Look up each word from text in Word Net**
- **If word is found, the set of synonyms from all Synsets are added to the query representation**
- **Weight each added word as .8 rather than 1.0**
- **Found results to be better than plain SMART**
 - Variable performance over queries
 - Major cause of error was when ambiguous words' Synsets are used in expansion

Use of Thesauri for expansion:

- **General thesauri such as Roget's or WordNet have not been shown conclusively to improve results:**
 - may sacrifice precision to recall
 - not domain specific
 - not sense disambiguated
- **But, a currently active field of R & D**

Disambiguation

- **Non-relevant documents may be retrieved because they contain the query term,**
 - but the wrong *sense* of the query term
- **Need good Word Sense Disambiguation**

Sample ambiguous query:

*I would like information about developments in low-risk **instruments**, especially those being offered by companies specializing in bonds.*

Human Sense Disambiguation

- **Sources of influence known from psycholinguistics research:**
 - local context
 - the sentence / query containing the ambiguous word restricts the interpretation of the ambiguous word

Sample ambiguous query:

*I would like information about developments in low-risk **instruments**, especially those being offered by companies specializing in **bonds**.*

Human Sense Disambiguation

- **Sources of influence known from psycholinguistics research:**
 - local context
 - the sentence / query containing the ambiguous word restricts the interpretation of the ambiguous word
 - domain knowledge
 - the fact that a text is concerned with a particular domain activates only the sense appropriate to that domain
 - frequency data
 - the frequency of each sense in general usage affects its accessibility to the mind

Machine Readable Lexical Sources

- **Multiple entries for polysemous words**
- *Instrument*
 - Medical
 - Financial
 - Dental
 - Musical
 - Hardware
 - Empirical experimentation
 - General

Machine Readable Lexical Sources

- **Senses are ranked by frequency of occurrence in usage:**
 1. Musical
 2. Hardware
 3. General
 4. Medical
 5. Dental
 6. Financial
 7. Empirical experimentation

Corpus-based Word Sense Disambiguation

- **Supervised learning from manually sense-tagged corpora**
 - allows development of algorithms which can correctly tag each word with its correct sense
 - utilizes context, which then proves essential in real-time disambiguation
 - usually a small window of words surrounding the ambiguous term
- **Issues**
 - time & cost in tagging the training sample
 - need to retag for new domains or genres

Word Sense Disambiguation

- **Impact on retrieval results**
 - Results vary
 - by approach used
 - by query (short queries, especially)
 - by engine
 - Some consider it a proven technique for improving Precision
 - Some are concerned about the trade-off in efficiency

Statistical Thesauri

- **Automatic thesaurus construction**
 - Classes of terms produced are not necessarily synonymous, nor broader, nor narrower
 - Rather, words that tend to co-occur with head term
 - Effectiveness varies considerably depending on technique used

Automatic Thesaurus Construction (Salton)

- **Document Collection Based**
 - based on index term similarities
 - compute vector similarities for each pair of documents
 - if sufficiently similar, create a thesaurus entry for each term which includes terms from similar document

Sample Automatic Thesaurus Entries:

408 dislocation 411 coercive
junction demagnetize
minority-carrier flux-
leakage
point contact hysteresis
recombine induct
transition insensitive
409 blast-cooled magnetoresistance
heat-flow square-loop
heat-transfer threshold
410 anneal 412 longitudinal
strain transverse

Dynamic Automatic Thesaurus Construction

- **Thesaurus short-cut**
 - Run at query time
 - Take all terms in query into consideration at once
 - Look at frequent words and phrases in top retrieved documents and add these to the query
 - = Automatic Relevance Feedback

Expansion by an Association Thesaurus

Query: *Impact of the 1986 Immigration Law*

Phrases retrieved by association in corpus

- *illegal immigration*
- *amnesty program*
- *immigration reform law*
- *editorial page article*
- *naturalization service*
- *civil fines*
- *new immigration law*
- *legal immigration law*
- *employer sanctions*
- *statutes*
- *applicability*
- *seeking amnesty*
- *legal status*
- *immigration act*
- *undocumented workers*
- *guest worker*
- *sweeping immigration*
- *undocumented aliens*

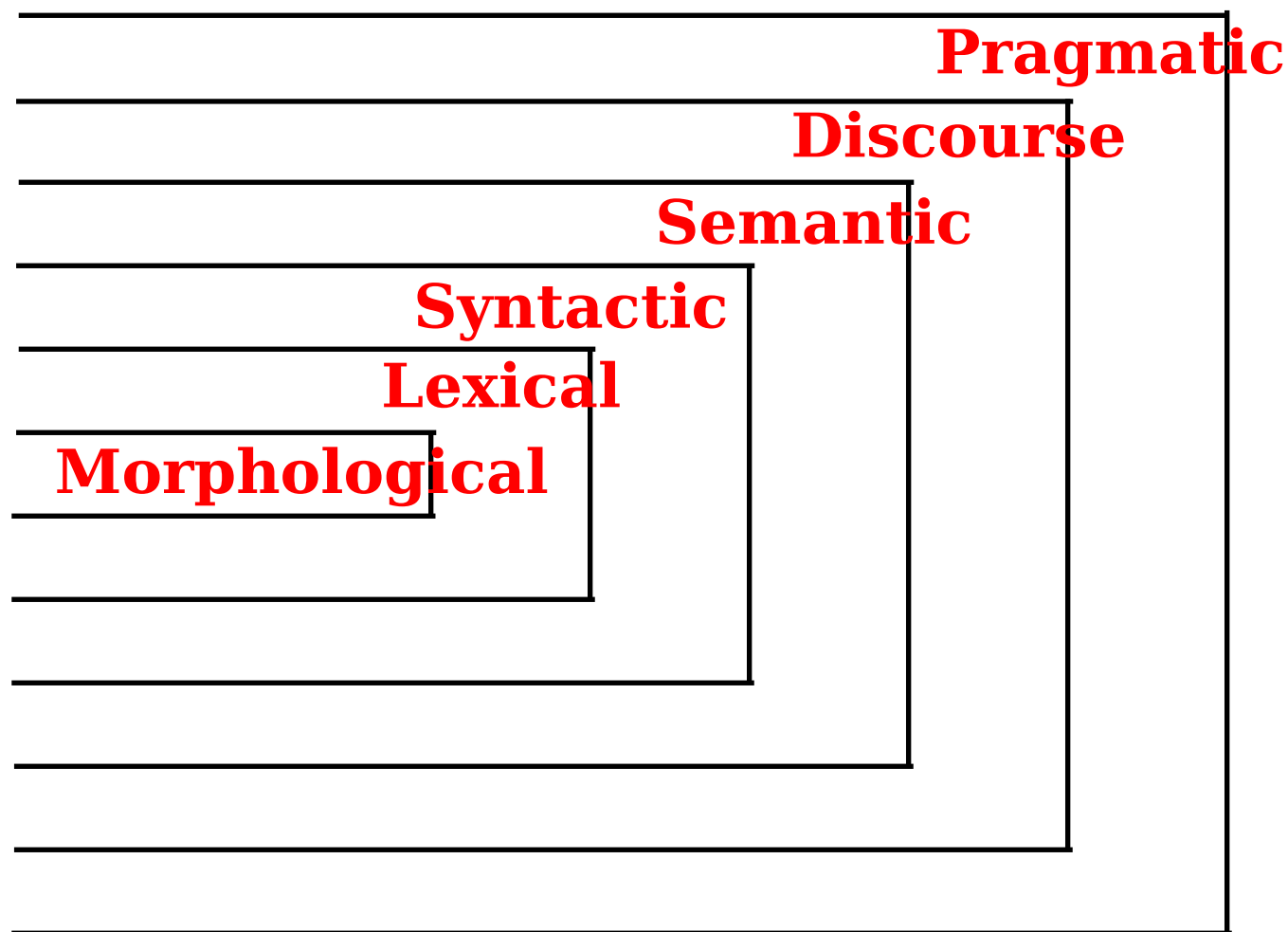
NLP-based Indexing

- the computational process of **identifying**, **selecting**, and **extracting** useful information from massive volumes of textual data:
 - for potential review by indexers
 - or stand-alone representation of content
 - using Natural Language Processing

Natural Language Processing

- **a range of computational techniques**
 - **for analyzing and representing naturally occurring texts**
 - **at one or more levels of linguistic analysis**
 - **for the purpose of achieving human-like language processing**
 - **for a range of tasks or applications**

Levels of Language Understanding



What can NLP Indexing do?

- **Phrase recognition**
- **Disambiguation**
- **Concept expansion**

In Summary:

- **There exist a range of approaches for representing documents and queries**
- **Each needs to be evaluated in terms of their ability to accomplish the goals of your application**
- **Web applications have opened a whole new world of possible variations on the traditional indexing approaches**